# Multi-view Domain Generalization for Visual Recognition

Li Niu[1], Wen Li[2], Dong Xu[3]
[1]Interdisciplinary Graduate School, Nanyang Technological University, Singapore
[2]Computer Vision Laboratory, ETH Zurich, Switzerland
[3]School of Computer Engineering, Nanyang Technological University, Singapore, and
School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia
lniu002@ntu.edu.sg, liwen@vision.ee.ethz.ch, dongxudongxu@gmail.com

## Abstract

*In this paper, we propose a new multi-view domain generalization (MVDG) approach for visual recognition, in which we aim to use the source domain samples with multiple types of features (*i.e.*, multi-view features) to learn robust classifiers that can generalize well to any unseen target domain. Considering the recent works show the domain generalization capability can be enhanced by fusing multiple SVM classifiers, we build upon exemplar SVMs to learn a set of SVM classifiers by using one positive sample and all negative samples in the source domain each time. When the source domain samples come from multiple latent domains, we expect the weight vectors of exemplar SVM classifiers can be organized into multiple hidden clusters. To exploit such cluster structure, we organize the weight vectors learnt on each view as a weight matrix and seek the low-rank representation by reconstructing this weight matrix using itself as the dictionary. To enforce the consistency of inherent cluster structures discovered from the weight matrices learnt on different views, we introduce a new regularizer to minimize the mismatch between any two representation matrices on different views. We also develop an efficient alternating optimization algorithm and further extend our MVDG approach for domain adaptation by exploiting the manifold structure of unlabeled target domain samples. Comprehensive experiments for visual recognition clearly demonstrate the effectiveness of our approaches for domain generalization and domain adaptation.*

## 1. Introduction

In many visual recognition tasks, the training and testing samples are with different data distributions. Recently, a large number of domain adaptation methods [20, 19, 1, 17, 6, 24, 14, 13, 15, 26, 27] have been proposed to explicitly cope with the data distribution mismatch between the training samples from the source domain and the testing samples from the target domain. Meanwhile, the domain generalization techniques [30, 38, 31] were also developed to learn robust classifiers that can generalize well to any unseen target domain. Please refer to Section 2 for the recent works on domain generalization and adaptation.

In most existing domain generalization/adaptation approaches, only one type of feature is used during the training and testing process. When multiple types of features are available for the training and testing samples, the visual recognition results can be improved by fusing multi-view features (See Section 2 for the related works on multi-view learning). Recently, multi-view domain adaptation methods [3, 41, 39] were also proposed to reduce the data distribution mismatch and simultaneously fuse multi-view features. In [3], Blitzer *et al.* proposed to learn the projection matrices by using Canonical Correlation Analysis (CCA) and adapt the source classifiers to the target domain based on the learnt projection matrices. In [41], the training samples are weighted similarly as in [24], while the prediction scores on different views are enforced to be close to each other. Yang and Gao proposed to add the Maximum Mean Discrepancy (MMD) based regularizer under the CCA framework in [39]. However, these multi-view domain adaptation methods [3, 41, 39] are only applicable when the target domain data is available.

In Section 3, we propose a multi-view domain generalization (MVDG) approach by using multi-view source domain samples to learn robust classifiers that can generalize well to any unseen target domain. Our method is motivated by the existing work [38] that experimentally demonstrates the domain generalization capability can be enhanced by fusing multiple SVM classifiers. Specifically, we build up our work on exemplar SVMs [29], in which we learn a set of SVM classifiers by using one positive sample and all negative samples in the source domain each time. As in [38, 18, 23], we also assume that the source domain samples come from multiple latent domains, so the weight vectors of exemplar SVM classifiers corresponding to the positive training samples from the same latent domain should be similar to each other, which means the weight vectors can be organized into multiple hidden clusters. To exploit such cluster structure, we first organize the weight vectors learnt on each view as a weight matrix, and then seek the low-rank representation (LRR) [28] for each weight matrix by reconstructing this matrix using itself as the dictionary. To enforce the consistency of inherent cluster structures discovered from the weight matrices learnt on different views, we introduce a new regularizer to minimize the mismatch between any two representation matrices on different views. We develop an efficient alternating optimization algorithm for our nontrivial optimization problem.

In Section 4, we further extend our MVDG to multi-view domain adaptation (MVDA), in which we introduce a smoothness based regularizer to exploit the manifold structure of unlabeled target domain samples. In section 5, we conduct comprehensive experiments and the results clearly demonstrate our approaches are better than related methods for visual recognition.

Our major contribution is an effective multi-view domain generalization method MVDG and its extended version MVDA. To the best of our knowledge, our work is the first to study the domain generalization problem under the multi-view setting.

## 2. Related Work

Our work is related to the domain generalization approaches [30, 38]. In [30], Muandet *et al*. proposed to reduce the marginal distribution mismatch between multiple latent domains while keeping their conditional distributions. But domain labels are required in [30], which are usually unavailable in many real-world applications. The most related work [38] aims to exploit the low-rank structure in positive training samples based on exemplar SVMs [29]. All the existing approaches [30, 38] focus on the single-view learning setting, while our work is the first for multi-view domain generalization.

In this work, our MVDG is extended for domain adaptation, so we also discuss the existing domain adaptation methods here. The existing domain adaptation methods can be roughly categorized into feature-based methods [20, 19, 1, 17], classifier-based methods [6, 14, 13, 15, 27], and instance-reweighting methods [24]. However, the above works did not discuss how to cope with multi-view source domain samples, which is the focus of our work. As discussed in Section 1, several domain adaptation methods [3, 41, 39] were recently proposed under the multi-view setting. However, they are only applicable when the target domain samples are available.

Our work is more related to recent latent domain discovering methods [18, 23]. The work in [23] uses the clustering method to partition the source domain samples into different latent domains, while the approach in [18] aims to maximize the separability of different latent domains based on the MMD criterion [24]. After discovering the latent domains, the classifiers trained for each latent domain are fused to predict the target domain samples. However, the number of latent domains is required and how to effectively utilize multi-view features was not discussed in the above methods.

Finally, we discuss the difference between our work and the existing multi-view learning methods [21, 16, 11]. The approach in [21] proposed to employ Kernel Canonical Correlation Analysis (KCCA) as the preprocessing step and then train SVM classifiers based on the transformed features. In [16], the above two-stage learning problem was formulated as a unified optimization problem. In [11], a low-rank common subspace was learnt among different views. Moreover, several multi-view semi-supervised learning methods [4, 34] were developed. For manifold regularization based methods, the Laplacian matrices from multi-view features are combined for semi-supervised learning in [34], while the semi-supervised Laplacian regularization is employed under the framework of KCCA in [2]. In co-training [4], the confident unlabeled training samples selected by using the classifier on one view are added into the labeled data set to learn the classifier on another view. However, the above multi-view learning methods assume the training and testing samples are from the same data distribution. In contrast, this assumption is not required in our multi-view domain generalization and adaptation approaches.

## 3. Multi-view Domain Generalization

In this section, we propose our multi-view domain generalization (MVDG) approach. For ease of presentation, a vector/matrix is denoted by a lowercase/uppercase letter in boldface. The transpose of a vector/matrix is denoted using the superscript $'$. We also denote $\mathbf{0}_n, \mathbf{1}_n \in \mathbb{R}^n$ as the $n$-dim column vectors of all zeros and all ones, respectively. When the dimensionality is obvious, we use $\mathbf{0}$ and $\mathbf{1}$ instead of $\mathbf{0}_n$ and $\mathbf{1}_n$. We also use $\mathbf{O}$ and $\mathbf{I}$ to denote the matrix of all zeros and identity matrix, respectively. Moreover, we denote $\mathbf{A} \circ \mathbf{B}$ as the element-wise product between two matrices. The inequality $\mathbf{a} \leq \mathbf{b}$ means that $a_i \leq b_i$ for $i = 1, \ldots, n$. We also denote $\mathbf{A}^{-1}$ as the inverse matrix of $\mathbf{A}$.

In this work, we study the multi-view domain generalization problem under the binary classification setting. Suppose we have $n$ positive training samples and $m$ negative training samples in the source domain, and each sample is represented as $V$ views of features. We denote each positive sample as $\mathbf{x}_i^+ = (\mathbf{x}_i^{1+}, \ldots, \mathbf{x}_i^{V+})$, $i = 1, \ldots, n$, and each negative sample as $\mathbf{x}_j^- = (\mathbf{x}_j^{1-}, \ldots, \mathbf{x}_j^{V-})$, $j = 1, \ldots, m$.

### 3.1. Domain Generalization with Exemplar SVMs

The key issue in domain generalization is to enhance the domain generalization capability of classifiers learnt from the training data. The recent works [18, 23] proposed to discover multiple latent domains from the training data, and train the discriminative classifiers for each latent domain. By fusing the classifiers from different latent domains, the integrated classifiers are robust to the changes of domain distributions, and thus generalize well for predicting the test samples from any unseen target domain.

However, the variance of training data in the real world applications may be affected by many hidden factors which usually overlap and interact with each other in complicated ways. As a result, discovering latent domains becomes a nontrivial task. Instead of explicitly discovering the latent domains, the recent work [38] proposed to exploit the intrinsic low-rank structure of positive training samples. In particular, their work builds upon the exemplar SVMs [29], in which multiple SVM classifiers are learnt by using one positive training sample and all negative training samples each time. Since those works [38, 29] were proposed for single-view training data (*i.e.*, $V = 1$), we omit the superscript $v$ in this section for ease of presentation. Let us denote $f_i(\mathbf{x}) = \mathbf{w}_i'\mathbf{x}$ as the exemplar SVM classifier learnt by using the $i$-th positive sample $\mathbf{x}_i^+$ and all negative samples[1] $\{\mathbf{x}_j^-|_{j=1}^m\}$. The objective of exemplar SVMs can be written as follows,

$$\min_{\mathbf{w}_i,\xi_i,\epsilon_{ij}} \quad \frac{1}{2}\sum_{i=1}^n \|\mathbf{w}_i\|^2 + C\sum_{i=1}^n \xi_i + C\sum_{i=1}^n\sum_{j=1}^m \epsilon_{ij} \quad (1)$$
$$\text{s.t.} \quad \mathbf{w}_i'\mathbf{x}_i^+ \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i,$$
$$\mathbf{w}_i'\mathbf{x}_j^- \leq -1 + \epsilon_{ij}, \quad \epsilon_{ij} \geq 0, \quad \forall i, \forall j,$$

where $\|\mathbf{w}_i\|^2$ is the regularization term, $\xi_i$'s and $\epsilon_{ij}$'s are the slack variables, and $C$ is the tradeoff parameter.

Since the positive training samples from the same latent domain are similar to each other, the prediction scores of positive samples from those exemplar classifiers should be low-rank. The work in [38] further employs a nuclear norm based regularizer on the prediction score matrix to exploit the intrinsic low-rank property of positive training samples. Nevertheless, this method only considers single-view training data. We show that when the training data consists of multi-view features, it is beneficial to jointly exploit the low-rank structures of the exemplar classifiers learnt on multiple views.

### 3.2. Multi-view Domain Generalization

When the training data consists of multi-view features, we learn exemplar SVMs on each view. Let us denote $f_i^v(\mathbf{x}^v) = \mathbf{w}_i^{v\prime}\mathbf{x}^v$ as the exemplar classifier on the $v$-th view learnt by using $\mathbf{x}_i^{v+}$ and $\{\mathbf{x}_j^{v-}|_{j=1}^m\}$ as the training data, and also denote $\mathbf{W}^v = [\mathbf{w}_1^v, \dots, \mathbf{w}_n^v]$ as the weight matrix of all exemplar classifiers on the $v$-th view.

The positive training samples may come from multiple latent domains, so the exemplar classifiers corresponding to the positive training samples from the same latent domain should be similar to each other, which means the weight vectors $\mathbf{w}_i^v$'s may come from multiple hidden clusters. In this work, we exploit such membership information (*i.e.*, which cluster each weight vector $\mathbf{w}_i^v$ belongs to) by using low-rank representation (LRR) [28], which has shown promising results in various real-world applications [37, 36]. Specifically, we seek a low-rank representation matrix $\mathbf{Z}^v \in \mathbb{R}^{n\times n}$ for each view such that the weight matrix can be represented as $\mathbf{W}^v = \mathbf{W}^v\mathbf{Z}^v + \mathbf{E}^v$, in which $\mathbf{E}^v$ is an error term and expected to be close to zeros. It has been shown in LRR that the representation matrix $\mathbf{Z}^v$ encodes the membership information of the samples, where the within-cluster entries of $\mathbf{Z}^v$ are usually dense, while the between-cluster entries of $\mathbf{Z}^v$ are usually sparse under certain conditions.

On one hand, with LRR, we expect to obtain a low-rank representation matrix $\mathbf{Z}^v$. By jointly learning the weight matrix $\mathbf{W}^v$ and the low-rank representation matrix $\mathbf{Z}^v$, we encourage $\mathbf{W}^v$ to exhibit clear cluster structure, namely, the weight vectors $\mathbf{w}_i^v$'s learnt by using the positive samples from the same latent domain should be similar to each other, while those from different latent domains are well distinguished.

On the other hand, when the training samples are with multi-view features, the membership information inferred based on $\mathbf{W}^v$'s on different views should be consistent. It is hard to directly enforce such consistency based on the weight matrices $\mathbf{W}^v$'s, since the weight matrices learnt using different views of features are in different feature spaces. Nevertheless, based on our low-rank representation, we can easily introduce the consistency by enforcing the the representation matrices $\mathbf{Z}^v$'s from different views to be close to each other with our newly proposed regularizer $\sum_{v,\tilde{v}:v\neq\tilde{v}} \|\mathbf{Z}^v - \mathbf{Z}^{\tilde{v}}\|_F^2$.

Based on the above discussions, we formulate our optimization problem as follows,

$$\min_{\substack{\mathbf{Z}^v,\mathbf{W}^v,\mathbf{E}^v\\ \xi_i^v,\epsilon_{ij}^v}} \sum_{v=1}^V \left(\frac{1}{2}\|\mathbf{W}^v\|_F^2 + C\sum_{i=1}^n \xi_i^v + C\sum_{i=1}^n\sum_{j=1}^m \epsilon_{ij}^v\right)$$
$$+\sum_{v=1}^V \left(\lambda_2\|\mathbf{E}^v\|_F^2 + \lambda_3\|\mathbf{Z}^v\|_*\right) + \frac{\gamma}{2}\sum_{v,\tilde{v}:v\neq\tilde{v}} \|\mathbf{Z}^v - \mathbf{Z}^{\tilde{v}}\|_F^2 \quad (2)$$
$$\text{s.t.} \quad \mathbf{w}_i^{v\prime}\mathbf{x}_i^{v+} \geq 1 - \xi_i^v, \quad \xi_i^v \geq 0, \quad \forall v, \forall i, \quad (3)$$
$$\mathbf{w}_i^{v\prime}\mathbf{x}_j^{v-} \leq -1 + \epsilon_{ij}^v, \quad \epsilon_{ij}^v \geq 0, \quad \forall v, \forall i, \forall j, \quad (4)$$
$$\mathbf{W}^v = \mathbf{W}^v\mathbf{Z}^v + \mathbf{E}^v, \quad \forall v, \quad (5)$$

[1] We do not explicitly use the bias term. Instead, we append 1 to each feature vector.

where $\|\mathbf{W}^v\|_F^2$ is the regularizer for exemplar classifiers, $\xi_i^v$, $\epsilon_i^v$ are the slack variables similarly as in (1), and $C$, $\lambda_2$, $\lambda_3$, and $\gamma$ are the trade-off parameters. The term $\|\mathbf{E}^v\|_F^2$ is used to enforce the error term $\mathbf{E}^v$ to be close to zeros, and the regularizer $\|\mathbf{Z}^v\|_*$ is the nuclear norm of $\mathbf{Z}^v$ and used to enforce $\mathbf{Z}^v$ to be low-rank.

### 3.3. Optimization

To solve the problem in (2), we first introduce an intermediate variable $\mathbf{G}^v$ for each $\mathbf{W}^v$, such that we employ low-rank representation (LRR) on $\mathbf{G}^v$ instead of $\mathbf{W}^v$ and enforce $\mathbf{G}^v$ to be close to $\mathbf{W}^v$. Specifically, we arrive at the new objective as follows,

$$\min_{\substack{\mathbf{Z}^v,\mathbf{W}^v,\mathbf{G}^v\\ \mathbf{E}^v,\xi_i^v,\epsilon_{ij}^v}} \sum_{v=1}^{V}\left(\frac{1}{2}\|\mathbf{W}^v\|_F^2 + C\sum_{i=1}^{n}\xi_i^v + C\sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{ij}^v\right)$$
$$+\sum_{v=1}^{V}\left(\lambda_1\|\mathbf{W}^v-\mathbf{G}^v\|_F^2+\lambda_2\|\mathbf{E}^v\|_F^2+\lambda_3\|\mathbf{Z}^v\|_*\right)$$
$$+\frac{\gamma}{2}\sum_{v,\tilde{v}:v\neq\tilde{v}}\|\mathbf{Z}^v-\mathbf{Z}^{\tilde{v}}\|_F^2 \quad (6)$$

$$\text{s.t.} \quad \mathbf{w}_i^{v\prime}\mathbf{x}_i^{v+} \geq 1-\xi_i^v, \quad \xi_i^v \geq 0, \quad \forall v, \forall i, \quad (7)$$
$$\mathbf{w}_i^{v\prime}\mathbf{x}_j^{v-} \leq -1+\epsilon_{ij}^v, \quad \epsilon_{ij}^v \geq 0, \quad \forall v, \forall i, \forall j, \quad (8)$$
$$\mathbf{G}^v = \mathbf{G}^v\mathbf{Z}^v + \mathbf{E}^v, \quad \forall v, \quad (9)$$

where $\lambda_1$ is the tradeoff parameter. It can be observed that the problem in (2) is a special case of (6) when $\lambda_1$ approaches $+\infty$. We solve (6) by using the alternating optimization approach. Specifically, we iteratively update two sets of variables $\{\mathbf{Z}^v, \mathbf{E}^v\}$ and $\{\mathbf{W}^v, \mathbf{G}^v, \xi_i^v, \epsilon_{ij}^v\}$ until the objective of (6) converges.

**Update $\mathbf{Z}^v$ and $\mathbf{E}^v$:** When fixing $\mathbf{W}^v$, $\mathbf{G}^v$, $\xi_i^v$, and $\epsilon_{ij}^v$, the problem in (6) reduces to the following problem,

$$\min_{\mathbf{Z}^v,\mathbf{E}^v} \sum_{v=1}^{V}(\lambda_2\|\mathbf{E}^v\|_F^2+\lambda_3\|\mathbf{Z}^v\|_*)+\frac{\gamma}{2}\sum_{v,\tilde{v}:v\neq\tilde{v}}\|\mathbf{Z}^v-\mathbf{Z}^{\tilde{v}}\|_F^2 \quad (10)$$
$$\text{s.t.}\ \mathbf{G}^v = \mathbf{G}^v\mathbf{Z}^v+\mathbf{E}^v, \quad \forall v. \quad (11)$$

By introducing the auxiliary variable $\mathbf{P}^v$ (*resp.*, $\mathbf{Q}^v$) to replace $\mathbf{Z}^v$ in $\|\mathbf{Z}^v\|_*$ (*resp.*, $\mathbf{Z}^v$ in the constraint (11)), the problem in (10) can be solved by using inexact augmented Lagrange Multiplier (ALM) method [5], which aims to minimize the following augmented Lagrangian function:

$$\mathcal{L} = \sum_{v=1}^{V}(\lambda_2\|\mathbf{E}^v\|_F^2+\lambda_3\|\mathbf{P}^v\|_*)+\frac{\gamma}{2}\sum_{v,\tilde{v}:v\neq\tilde{v}}\|\mathbf{Z}^v-\mathbf{Z}^{\tilde{v}}\|_F^2 \quad (12)$$
$$+\sum_{v=1}^{V}\langle\mathbf{S}^v,\mathbf{Z}^v-\mathbf{P}^v\rangle+\sum_{v=1}^{V}\langle\mathbf{T}^v,\mathbf{Z}^v-\mathbf{Q}^v\rangle$$
$$+\sum_{v=1}^{V}\langle\mathbf{R}^v,\mathbf{G}^v-\mathbf{G}^v\mathbf{Q}^v-\mathbf{E}^v\rangle+\frac{\mu}{2}\sum_{v=1}^{V}\|\mathbf{Z}^v-\mathbf{P}^v\|_F^2$$
$$+\frac{\mu}{2}\sum_{v=1}^{V}\|\mathbf{Z}^v-\mathbf{Q}^v\|_F^2+\frac{\mu}{2}\sum_{v=1}^{V}\|\mathbf{G}^v-\mathbf{G}^v\mathbf{Q}^v-\mathbf{E}^v\|_F^2,$$

where $\mathbf{S}^v$, $\mathbf{T}^v$, and $\mathbf{R}^v$ are the Lagrangian multipliers, $\mu > 0$ is a penalty parameter. Basically, the inexact ALM method is to iteratively update each variable in the augmented Lagrangian function (12) until the termination criterion is satisfied. We list the steps to solve (12) in Algorithm 1. In particular, the problem in (22) can be solved by using the Singular Value Threshold (SVT) method [7], similarly as in LRR. The equation in (23) (*resp.*, (25)) can be obtained by setting the derivative of (12) w.r.t. $\mathbf{Q}^v$ (*resp.*, $\mathbf{E}^v$) to zeros. The problem in (24) can be solved by stacking the vectorizations of all $\mathbf{Z}^v$'s to $\bar{\mathbf{Z}} \in \mathbb{R}^{V\times n^2}$ and setting the derivative of (24) w.r.t. $\bar{\mathbf{Z}}$ to zeros.

**Update $\mathbf{W}^v$, $\mathbf{G}^v$, $\xi_i^v$, $\epsilon_{ij}^v$:** When fixing $\mathbf{Z}^v$ and equivalently replacing $\mathbf{E}^v$ with $\mathbf{G}^v-\mathbf{G}^v\mathbf{Z}^v$, the problem in (6) reduces to the following problem,

$$\min_{\substack{\mathbf{W}^v,\mathbf{G}^v\\ \xi_i^v,\epsilon_{ij}^v}} \sum_{v=1}^{V}(\frac{1}{2}\|\mathbf{W}^v\|_F^2+C\sum_{i=1}^{n}\xi_i^v+C\sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{ij}^v \quad (13)$$
$$+\lambda_1\|\mathbf{W}^v-\mathbf{G}^v\|_F^2+\lambda_2\|\mathbf{G}^v-\mathbf{G}^v\mathbf{Z}^v\|_F^2)$$
$$\text{s.t.}\ \mathbf{w}_i^{v\prime}\mathbf{x}_i^{v+} \geq 1-\xi_i^v, \quad \xi_i^v \geq 0, \quad \forall v, \forall i, \quad (14)$$
$$\mathbf{w}_i^{v\prime}\mathbf{x}_j^{v-} \leq -1+\epsilon_{ij}^v, \quad \epsilon_{ij}^v \geq 0, \quad \forall v, \forall i, \forall j. \quad (15)$$

The above problem can be separated into $V$ independent subproblems with one for each view. We solve the subproblem on each view by updating two sets of variables $\{\mathbf{W}^v$, $\xi_i^v$, $\epsilon_{ij}^v\}$ and $\mathbf{G}^v$ alternatively until the objective of (13) converges. Specifically, when fixing $\mathbf{G}^v$, we solve $\mathbf{W}^v$, $\xi_i^v$, and $\epsilon_{ij}^v$ by separately solving $n$ independent subproblems, in which each subproblem is related to one exemplar classifier. The subproblem w.r.t. the $i$-th exemplar classifier can be written as follows,

$$\min_{\mathbf{w}_i^v,\xi_i^v,\epsilon_{ij}^v} \frac{1}{2}\|\mathbf{w}_i^v\|^2+C(\xi_i^v+\sum_{j=1}^{m}\epsilon_{ij}^v)+\lambda_1\|\mathbf{w}_i^v-\mathbf{g}_i^v\|^2 \quad (16)$$
$$\text{s.t.}\ \mathbf{w}_i^{v\prime}\mathbf{x}_i^{v+} \geq 1-\xi_i^v, \quad \xi_i^v \geq 0, \quad (17)$$
$$\mathbf{w}_i^{v\prime}\mathbf{x}_j^{v-} \leq -1+\epsilon_{ij}^v, \quad \epsilon_{ij}^v \geq 0, \quad \forall j, \quad (18)$$

where $\mathbf{g}_i^v$ is the $i$-th column vector of $\mathbf{G}^v$. By respectively introducing the dual variables $\{\alpha_i^+$, $\beta_i^+\}$ and $\{\alpha_j^-$, $\beta_j^-\}$'s for the constraints in (17) and (18), we arrive at the dual form of (16) as follows,

$$\min_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}'\frac{\bar{\mathbf{X}}_i^{v\prime}\bar{\mathbf{X}}_i^v\circ(\mathbf{y}\mathbf{y}')}{2(1+2\lambda_1)}\boldsymbol{\alpha}+[\frac{2\lambda_1(\bar{\mathbf{X}}_i^{v\prime}\mathbf{g}_i^v)\circ\mathbf{y}}{1+2\lambda_1}-\mathbf{1}]'\boldsymbol{\alpha} \quad (19)$$
$$\text{s.t.} \quad \mathbf{0}\leq\boldsymbol{\alpha}\leq C\mathbf{1},$$

where $\bar{\mathbf{X}}_i^v = [\mathbf{x}_i^{v+}, \mathbf{x}_1^{v-}, \ldots, \mathbf{x}_m^{v-}]$, $\boldsymbol{\alpha} = [\alpha_i^+, \alpha_1^-, \ldots, \alpha_m^-]'$ and $\mathbf{y} = [1, -\mathbf{1}_m{}']'$. The problem in (19) is a quadratic programming (QP) problem, which can be solved by using the existing QP solvers. In our work, we develop an efficient SMO approach to solve (19) by updating one selected dual variable in each iteration. After obtaining $\boldsymbol{\alpha}$, we can recover $\mathbf{w}_i^v$ as follows,

$$\mathbf{w}_i^v = \frac{1}{1+2\lambda_1}(2\lambda_1\mathbf{g}_i^v+\bar{\mathbf{X}}_i^v(\mathbf{y}\circ\boldsymbol{\alpha})). \quad (20)$$

**Algorithm 1** Solving (12) with inexact ALM

```
Input: $\mathbf{G}^v, \lambda_2, \lambda_3, \gamma$
Initialize $\mathbf{Z}^v = \mathbf{E}^v = \mathbf{S}^v = \mathbf{T}^v = \mathbf{R}^v = \mathbf{O}$, $\rho = 0.1$,
    $\mu = 0.1$, $\mu_{max} = 10^6$, $\nu = 10^{-5}$, $N_{iter} = 10^6$.
for $t = 1 : N_{iter}$ do
    For $v = 1, \ldots, V$, update $\mathbf{P}^v$ by solving
        $\mathbf{P}^v = \arg\min_{\mathbf{P}^v} \lambda_3 \|\mathbf{P}^v\|_* + \frac{\mu}{2}\|\mathbf{P}^v - (\mathbf{Z}^v + \frac{\mathbf{S}^v}{\mu})\|_F^2. \quad (22)$
    For $v = 1, \ldots, V$, update $\mathbf{Q}^v$ by
        $\mathbf{Q}^v = (\mathbf{I} + \mathbf{G}^{v\prime}\mathbf{G}^v)^{-1}(\mathbf{G}^{v\prime}(\mathbf{G}^v - \mathbf{E}^v + \frac{\mathbf{R}^v}{\mu}) + \mathbf{Z}^v + \frac{\mathbf{T}^v}{\mu}). \quad (23)$
    For $v = 1, \ldots, V$, update $\mathbf{Z}^v$ by solving
        $\min_{\mathbf{Z}^v} \frac{\gamma}{2} \sum_{v,\tilde{v}:v \neq \tilde{v}} \|\mathbf{Z}^v - \mathbf{Z}^{\tilde{v}}\|_F^2 + \sum_{v=1}^{V} \mu \|\mathbf{Z}^v - \mathbf{H}^v\|_F^2, \quad (24)$
        where $\mathbf{H}^v = \frac{1}{2}(\mathbf{P}^v + \mathbf{Q}^v - \frac{1}{\mu}(\mathbf{S}^v + \mathbf{T}^v))$.
    For $v = 1, \ldots, V$, update $\mathbf{E}^v$ by
        $\mathbf{E}^v = \frac{\mu(\mathbf{G}^v - \mathbf{G}^v\mathbf{Q}^v) + \mathbf{R}^v}{2\lambda_2 + \mu}. \quad (25)$
    For $v = 1, \ldots, V$, update $\mathbf{S}^v$, $\mathbf{T}^v$, and $\mathbf{R}^v$ by
        $\mathbf{S}^v = \mathbf{S}^v + \mu(\mathbf{Z}^v - \mathbf{P}^v), \quad (26)$
        $\mathbf{T}^v = \mathbf{T}^v + \mu(\mathbf{Z}^v - \mathbf{Q}^v), \quad (27)$
        $\mathbf{R}^v = \mathbf{R}^v + \mu(\mathbf{G}^v - \mathbf{G}^v\mathbf{Q}^v - \mathbf{E}^v). \quad (28)$
    Update the parameter $\mu$ by $\mu = \min(\mu_{max}, (1+\rho)\mu)$.
    Break if $\|\mathbf{G}^v - \mathbf{G}^v\mathbf{Q}^v - \mathbf{E}^v\|_\infty < \nu$, $\|\mathbf{Z}^v - \mathbf{P}^v\|_\infty < \nu$,
        $\|\mathbf{Z}^v - \mathbf{Q}^v\|_\infty < \nu, \quad \forall v$.
end for
Output: $\mathbf{Z}^v$.
```

When fixing $\mathbf{W}^v$, $\xi_i^v$, and $\epsilon_{ij}^v$, we set the derivative of the objective function in (13) w.r.t. $\mathbf{G}^v$ to zeros, and then update $\mathbf{G}^v$ by using the closed-form solution as follows,

$$\mathbf{G}^v = \lambda_1 \mathbf{W}^v \left(\lambda_2(\mathbf{I} - \mathbf{Z}^v)(\mathbf{I} - \mathbf{Z}^v)' + \lambda_1 \mathbf{I}\right)^{-1}. \quad (21)$$

The whole algorithm is summarized in Algorithm 2.

In the testing stage, it is more reasonable to utilize the exemplar classifiers from the latent source domain which is close to the target domain. Inspired by [38], given a test sample, we fuse the exemplar classifiers which output higher prediction scores on this test sample. Formally, given a test sample $\mathbf{u} = (\mathbf{u}^1, \ldots, \mathbf{u}^V)$ with $\mathbf{u}^v$ denoting the $v$-th view feature, the prediction score of $\mathbf{u}$ can be obtained as follows,

$$f(\mathbf{u}) = \frac{1}{V} \sum_{v=1}^{V} \frac{1}{|\Gamma(\mathbf{u}^v)|} \sum_{i:i \in \Gamma(\mathbf{u}^v)} f_i^v(\mathbf{u}^v), \quad (29)$$

where $f_i^v(\mathbf{u}^v)$ is the prediction score obtained by applying the exemplar classifier $\mathbf{w}_i^v$ on $\mathbf{u}^v$, and $\Gamma(\mathbf{u}^v)$ denotes the index set of exemplar classifiers that output the top prediction scores on $\mathbf{u}^v$. In our experiments, we set the cardinality of $\Gamma(\mathbf{u}^v)$ (*i.e.*, $|\Gamma(\mathbf{u}^v)|$) as 5, as suggested in [38]. When predicting each test sample by using the exemplar classifiers with higher prediction scores, we conjecture this test sample is likely to be sampled from the most relevant latent source domain, from which the corresponding positive training samples are used to learn those selected exemplar classifiers. As a result, the fused classifier by using (29) can generalize well to any unseen target domain.

**Algorithm 2** Multi-view Domain Generalization (MVDG)

```
Input: Training data $\{\mathbf{x}_i^{v+}|_{i=1}^{n}\}$ and $\{\mathbf{x}_j^{v-}|_{j=1}^{m}\}$ with $V$
    views.
Initialize[2] $\mathbf{G}^v$'s.
repeat
    Update $\mathbf{Z}^v$'s by using Algorithm 1.
    repeat
        Update $\mathbf{W}^v$ by solving $n$ subproblems in the dual
            form (19) and recovering $\mathbf{W}^v$ by using (20) on
            each view.
        Update $\mathbf{G}^v$ by using (21) on each view.
    until The objective of (13) converges.
until The objective of (6) converges.
Output: The learnt classifier $\mathbf{W}^v$'s.
```

## 4. Extending MVDG for Domain Adaptation

When the unlabeled samples from the target domain are available during the training process, we extend our MVDG approach to multi-view domain adaptation (MVDA) by using those unlabeled samples for domain adaptation. In particular, we additionally utilize a Laplacian regularizer, such that the prediction scores using the exemplar classifiers should satisfy the smoothness constraint on the unlabeled samples in the target domain. We formulate our MVDA approach as follows,

$$\min_{\substack{\mathbf{Z}^v, \mathbf{W}^v, \mathbf{G}^v \\ \mathbf{E}^v, \xi_i^v, \epsilon_{ij}^v}} \sum_{v=1}^{V} \Big(\frac{1}{2}\|\mathbf{W}^v\|_F^2 + C\sum_{i=1}^{n} \xi_i^v + C\sum_{i=1}^{n}\sum_{j=1}^{m} \epsilon_{ij}^v$$

$$+\lambda_1\|\mathbf{W}^v - \mathbf{G}^v\|_F^2 + \lambda_2\|\mathbf{E}^v\|_F^2 + \lambda_3\|\mathbf{Z}^v\|_*\Big)$$

$$+\frac{\gamma}{2}\sum_{v,\tilde{v}:v\neq\tilde{v}} \|\mathbf{Z}^v - \mathbf{Z}^{\tilde{v}}\|_F^2 + \theta\sum_{v=1}^{V} \Omega(\mathbf{W}^v, \mathbf{L}^v, \mathbf{U}^v) \quad (30)$$

$$\text{s.t.} \quad \mathbf{w}_i^{v\prime}\mathbf{x}_i^{v+} \geq 1 - \xi_i^v, \quad \xi_i^v \geq 0, \quad \forall v, \forall i, \quad (31)$$

$$\mathbf{w}_i^{v\prime}\mathbf{x}_j^{v-} \leq -1 + \epsilon_{ij}^v, \quad \epsilon_{ij}^v \geq 0, \quad \forall v, \forall i, \forall j, \quad (32)$$

$$\mathbf{G}^v = \mathbf{G}^v\mathbf{Z}^v + \mathbf{E}^v, \quad \forall v, \quad (33)$$

where $\Omega(\mathbf{W}^v, \mathbf{L}^v, \mathbf{U}^v) = \text{tr}(\mathbf{W}^{v\prime}\mathbf{U}^v\mathbf{L}^v\mathbf{U}^{v\prime}\mathbf{W}^v)$ is the Laplacian regularizer, in which $\mathbf{L}^v$ is the Laplacian matrix for the target domain samples on the $v$-th view, $\mathbf{U}^v = [\mathbf{u}_1^v, \ldots, \mathbf{u}_N^v]$ is the target domain samples with $N$ being the

[2] We initialize $\mathbf{G}^v$ with its $i$-th column vector being the weight vector of exemplar classifiers trained by using the $i$-th positive training sample and all negative training samples on the $v$-th view.

total number of unlabeled target domain samples and $\mathbf{u}_i^v$ being the $v$-th view feature of the $i$-th sample. We construct Laplacian matrices $\mathbf{L}^v$'s based on cosine distance.

The solution to (30) can be similarly derived as that to (6). The only difference is that we respectively replace $\frac{1}{1+2\lambda_1}(\bar{\mathbf{X}}_i^{v\prime}\bar{\mathbf{X}}_i^v)$ and $\frac{1}{1+2\lambda_1}(\bar{\mathbf{X}}_i^{v\prime}\mathbf{g}_i^v)$ in the first and second terms with $\bar{\mathbf{X}}_i^{v\prime}(\mathbf{J}^v)^{-1}\bar{\mathbf{X}}_i^v$ and $\bar{\mathbf{X}}_i^{v\prime}(\mathbf{J}^v)^{-1}\mathbf{g}_i^v$ when solving (19), where $\mathbf{J}^v = (1+2\lambda_1)\mathbf{I} + 2\theta\mathbf{U}^v\mathbf{L}^v\mathbf{U}^{v\prime}$.

# 5. Experiments

In this section, we demonstrate the effectiveness of our proposed approaches for human action recognition and object recognition.

## 5.1. Datasets and Experimental Settings

We evaluate different methods using two human action datasets ACT4$^2$ [9] and Online RGBD Action Dataset (ORGBD) [40] as well as one object recognition dataset Office-Caltech [33, 19]. For performance evaluation, we report the recognition accuracy for all methods.

**ACT4$^2$ dataset:** The ACT4$^2$ dataset contains both RGB and depth videos from 14 representative classes of daily actions, which are captured by using different Kinect cameras from 4 viewpoints. Similarly as in [38], we treat the videos captured by each camera as one domain. Then, we mix the samples from several domains as the source domain, and use the remaining samples as the target domain. As suggested in [9], we use the samples from 2 cameras for training and the samples from the remaining 2 cameras for testing. So we have 6 settings in total.

**ORGBD dataset:** The Online RGBD Action Dataset (ORGBD) [40] provides the RGB-D videos from 7 types of actions. The whole dataset consists of 4 sets, among which the first three sets can be used for cross-environment action recognition as suggested in [40]. Set 1 and set 2 are recorded in the same environment while Set 3 is recorded in another environment. In our experiments, we combine two sets from different environments as training data and the remaining one as test data, which leads to 2 settings.

We use two-view features (*i.e.*, the RGB features and depth features) in our experiments. Specifically, we extract the improved dense trajectory features [35] from each RGB and depth video in both ACT4$^2$ and ORGBD datasets. Following [8], we obtain a 6000-dim feature vector for each RGB and depth video by concatenating the bag-of-words (BoW) features from three types of descriptors, in which each type of descriptors are encoded to the 2000-dim BoW feature.

**Office-Caltech dataset:** The Office-Caltech dataset [19] contains the images from four domains: Amazon (A), Caltech-256 (C), Digital SLR (D), and Webcam (W). Following [18, 38], we evaluate all methods using the 10 common categories among the 4 domains based on three settings, in which A and C (*resp.*, D and W; C, D, and W) are merged as the source domain while the remaining domains are merged as the target domain. For each image, we extract the DeCAF$_6$ feature [12] and the Caffe$_6$ [25] feature as two views of features.

## 5.2. Results of MVDG and MVDA

We first evaluate our proposed MVDG and MVDA approaches. In order to show the effectiveness of our multi-view learning approach, we report the results of a special case of MVDG, which is named MVDG (w/o co-reg), in which we remove the co-regularizer $\sum_{v,\tilde{v}:v\neq\tilde{v}} \|\mathbf{Z}^v - \mathbf{Z}^{\tilde{v}}\|_F^2$ in (6) by setting $\gamma$ to 0. Furthermore, to investigate the benefits after fusing multi-view features, we also report the results of our MVDG method by only utilizing one view of features, which are referred to as MVDG (RGB/DeCAF) and MVDG (depth/Caffe). Moreover, we include SVM [10] and Exemplar-SVM (ESVM) [29] as two baselines for comparison. For SVM, we train the SVM classifier on each view, and fuse the prediction scores on two views. For ESVM, we train one exemplar classifier for each positive sample on each view, and then employ the same prediction strategy as our MVDG method (see (29)). For our methods, we empirically fix the parameters as $C = 0.1$, $\lambda_1 = 100$, $\lambda_2 = 10$, $\lambda_3 = 0.1$, $\gamma = 100$, $\theta = 10^{-5}$ for all the settings on all datasets. For SVM and ESVM, we choose the optimal parameters according to their best accuracies on the test set.

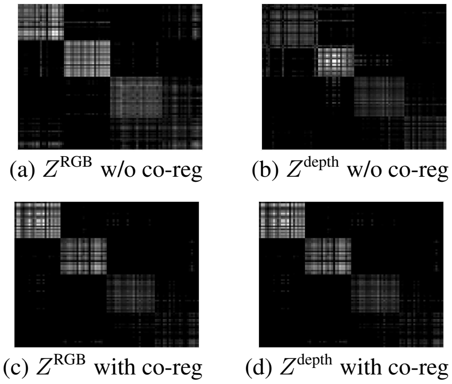
The experimental results are summarized in Table 1. We observe that ESVM is generally better than SVM, which shows that it is beneficial to fuse multiple exemplar classifiers for enhancing the domain generalization capacity. Note ESVM can be considered as a special case of our MVDG (w/o co-reg) without using LRR when learning the classifiers on each view. We also observe that MVDG (w/o co-reg) achieves better results than ESVM, which demonstrates it is beneficial to exploit the cluster structure of representation matrices by using LRR for domain generalization. It can also be observed that our special case MVDG (w/o co-reg) achieves better results than both MVDG (RGB/DeCAF) and MVDG (depth/Caffe), which shows it is helpful to fuse multi-view features.

Our MVDG method outperforms its special case MVDG (w/o co-reg), which validates the effectiveness of our newly proposed co-regularizer $\sum_{v,\tilde{v}:v\neq\tilde{v}} \|\mathbf{Z}^v - \mathbf{Z}^{\tilde{v}}\|_F^2$. So it is beneficial to exploit the cluster structure discovered from representation matrices on different views. Finally, our domain adaptation method MVDA further achieves better results than MVDG, which shows the effectiveness of MVDA for exploiting the unlabeled target domain samples in the training process based on the Laplacian regularizer.

We also give a visual comparison of the representation matrices $\mathbf{Z}^v$'s ($\mathbf{Z}^1$ and $\mathbf{Z}^2$ are denoted as $\mathbf{Z}^{\text{RGB}}$ and $\mathbf{Z}^{\text{depth}}$ respectively for better presentation) learnt by using

Table 1: Accuracies (%) of our proposed methods for human action recognition and object recognition. For comparison, we also report the results of the special cases of our MVDG method and the baseline methods SVM and ESVM.

| Dataset | $\text{ACT4}^2$ | | | | | | ORGBD | | Office-Caltech | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | 1,2 | 1,3 | 1,4 | 2,3 | 2,4 | 3,4 | 1,3 | 2,3 | A,C | D,W | C,D,W |
| Target | 3,4 | 2,4 | 2,3 | 1,4 | 1,3 | 1,2 | 2 | 1 | D,W | A,C | A |
| SVM [10] | 63.75 | 62.61 | 42.90 | 45.92 | 59.52 | 63.67 | 51.79 | 45.54 | 83.19 | 79.14 | 91.23 |
| ESVM [29] | 66.01 | 63.07 | 46.98 | 45.24 | 65.11 | 65.18 | 54.46 | 49.11 | 85.18 | 81.60 | 91.65 |
| MVDG (RGB/DeCAF) | 58.01 | 69.03 | 49.55 | 49.32 | 62.08 | 63.75 | 55.36 | 48.21 | 84.29 | 81.07 | 93.01 |
| MVDG (depth/Caffe) | 65.41 | 67.37 | 42.15 | 43.13 | 62.39 | 64.50 | 55.71 | 49.46 | 86.73 | 82.27 | 92.69 |
| MVDG (w/o co-reg) | 70.09 | 73.19 | 50.83 | 50.91 | 68.13 | 71.53 | 56.25 | 51.79 | 86.95 | 83.09 | 93.11 |
| MVDG | **72.81** | **76.21** | **52.04** | **53.17** | **69.79** | **72.96** | **58.04** | **53.57** | **87.62** | **83.47** | **93.32** |
| MVDA | **75.15** | **77.72** | **53.93** | **57.63** | **73.04** | **74.55** | **59.82** | **56.25** | **93.81** | **85.25** | **94.05** |



(a) $Z^{\text{RGB}}$ w/o co-reg (b) $Z^{\text{depth}}$ w/o co-reg

(c) $Z^{\text{RGB}}$ with co-reg (d) $Z^{\text{depth}}$ with co-reg

Figure 1: Illustration of the representation matrices $\mathbf{Z}^v$'s for the action "Put On" on the $\text{ACT4}^2$ dataset under the setting $1, 4 \rightarrow 2, 3$.

our MVDG and its special case MVDG (w/o co-reg) in Figure 1. Recall that the representation matrix $\mathbf{Z}^v$ encodes the membership information of positive training samples, in which the within-cluster (*resp.*, between-cluster) entities are usually dense (*resp.*, sparse), so $\mathbf{Z}^v$ is expected to be block-diagonal in ideal cases. It can be observed all four representation matrices are near block-diagonal with each block corresponding to one latent domain, which demonstrates the effectiveness of employing LRR for discovering latent domains on each view. It is worth mentioning that we only mix the training samples from two domains (i.e., camera 1 and camera 4) as the source domain samples, but there are actually 4 latent domains discovered by using our methods (See Figure 1). One possible explanation is that for the action "Put On", actors may put on clothes from the opposite directions, which may lead to two latent domains for the videos captured by each camera. After employing our co-regularizer, we also observe the two representation matrices learnt by using MVDG (the bottom row) are more consistent and exhibit relatively better block-diagonal structure when compared with those learnt by using MVDG (w/o co-reg) (the top row). The result indicates it is beneficial to use our newly proposed co-regularizer to exploit the intrinsic cluster structure of latent domains based on multiple views of features.

## 5.3. Comparison with the State-of-the-art

We compare our methods with the state-of-the-art methods for domain generalization and domain adaptation. We use the same parameter setting for our methods as in Section 5.2. For the baselines methods, we choose the optimal parameters according to their best accuracies on the test set. Due to space limitation, we only report the mean accuracy over the 6 (*resp.*, 2, 3) settings for the $\text{ACT4}^2$ (*resp.*, ORGBD, Office-Caltech) dataset.

### 5.3.1 Domain Generalization

**Baselines:** We compare our MVDG method with three groups of baselines: the multi-view learning methods, the domain generalization methods, and the latent domain discovering methods. The multi-view learning methods include KCCA [21], SVM-2K [16] and low-rank common subspace (LRCS) [11], which treat $\text{RGB/DeCAF}_6$ features and $\text{depth/Caffe}_6$ features as two views.

The domain generalization methods contain the domain-invariant component analysis (DICA) method [30] and the low-rank exemplar SVM (LRESVM) method [38]. We employ DICA and LRESVM on each view and then use the late-fusion strategy to fuse the prediction scores from two views.

To discover the latent domains, we use the approaches in [23, 18]. We train the classifiers for each latent domain, and then apply two strategies named "match" and "ensemble" for testing as suggested in [38]. As sub-categorization is similar with latent domain discovery and applicable in our task, we additionally compare our work with the discriminative sub-categorization(Sub-Cate) method [22]. For all the methods mentioned above, we apply them on each view independently and employ the late-fusion strategy to fuse the prediction scores from two views.

**Experimental Results:** The experimental results are summarized in Table 2. Multi-view learning methods LRCS, SVM-2K and KCCA outperform SVM because they can better exploit two-view features.

Sub-Cate and the domain generalization methods DICA and LRESVM generally achieve better results than SVM, which shows the advantage of exploiting the intrinsic structure when using training data from a mixture of latent do-

Table 2: Mean accuracies (%) of different methods over multiple settings on each dataset without using target domain data in the training process. The best results are denoted in boldface.

| Dataset | ACT4$^2$ | ORGBD | Office |
|---|---|---|---|
| SVM [10] | 56.40 | 48.67 | 84.52 |
| ESVM [29] | 58.60 | 51.79 | 86.14 |
| LRCS [11] | 59.72 | 52.68 | 85.28 |
| SVM-2K [16] | 59.68 | 50.00 | 86.10 |
| KCCA [21] | 57.72 | 51.34 | 86.33 |
| DICA [30] | 59.10 | 47.32 | 86.12 |
| LRESVM [38] | 62.61 | 53.57 | 87.04 |
| [18](match) | 57.83 | 50.00 | 86.47 |
| [18](ensemble) | 58.42 | 51.79 | 86.06 |
| [23](match) | 55.21 | 44.65 | 85.75 |
| [23](ensemble) | 57.78 | 50.45 | 84.81 |
| Sub-Cate [22] | 59.71 | 52.68 | 86.64 |
| MVDG | **66.16** | **55.81** | **88.13** |

mains. For the latent domain discovering methods [23, 18], the results obtained by using the "ensemble" strategy are better than SVM, which demonstrates it is effective to discover the latent domains. The results using the "ensemble" strategy are also generally better when compared with the "match" strategy.

Finally, our MVDG method outperforms all the baselines on all datasets, which demonstrates our MVDG method can enhance the domain generalization capability and effectively fuse the multi-view features simultaneously.

### 5.3.2 Domain Adaptation

For domain adaptation, we further utilize the unlabeled target domain samples during the training process.

**Baselines:** We compare our MVDA method with three groups of baselines: the domain adaptation methods, the multi-view semi-supervised learning methods, and the existing multi-view domain adaptation methods. The domain adaptation baseline methods include DASVM [6], K-MM [24], TCA [32], SA [17], DIP [1], GFK [19], and S-GF [20]. We apply the above domain adaptation methods on each view and use the late fusion strategy to fuse the prediction scores from two views.

We compare our MVDA method with multi-view semi-supervised learning methods Co-training [4] and Co-LapSVM [34], as well as the existing multi-view domain adaptation methods Coupled [3], MVTL_LM [41], and MDT [39], which fuse the multi-view features and simultaneously reduce the domain distribution mismatch. We further compare our MVDA with LRCS [11] by using the target samples as the dictionary as suggested in [11].

**Experimental Results:** The experimental results are summarized in Table 3. We also add the results of SVM from Table 2 for comparison. We observe that the domain adaptation methods DASVM, KMM, TCA, SA, DIP, GFK, and S-GF generally achieve better results than SVM, which shows it is beneficial to reduce the domain distribution mismatch between the source domain and the target domain. The multi-view semi-supervised learning methods Co-training and Co-LapSVM outperform SVM, which demonstrates it is helpful to utilize the unlabeled target domain data.

Table 3: Mean accuracies (%) of different methods over multiple settings on each dataset after using target domain data in the training process. The best results are denoted in boldface.

| Dataset | ACT4$^2$ | ORGBD | Office |
|---|---|---|---|
| SVM [10] | 56.40 | 48.66 | 84.52 |
| DASVM [6] | 60.22 | 50.45 | 85.60 |
| KMM [24] | 59.46 | 52.12 | 86.34 |
| TCA [32] | 59.12 | 48.66 | 85.79 |
| SA [17] | 63.42 | 52.24 | 86.79 |
| DIP [1] | 58.86 | 54.46 | 86.58 |
| GFK [19] | 60.61 | 53.13 | 86.22 |
| SGF [20] | 56.17 | 52.23 | 85.78 |
| Co-training [4] | 62.15 | 53.13 | 87.96 |
| Co-LapSVM [34] | 61.57 | 52.68 | 88.20 |
| Coupled [3] | 64.79 | 54.02 | 86.48 |
| MVTL_LM [41] | 63.70 | 55.36 | 87.76 |
| MDT [39] | 64.97 | 54.46 | 86.87 |
| LRCS [11] | 62.07 | 55.81 | 86.12 |
| MVDA | **68.67** | **58.04** | **91.04** |

For the multi-view domain adaptation methods, Coupled, MVTL_LM, and MDT outperform the multi-view learning methods reported in Table 2, possibly because they further consider the domain distribution mismatch. When compared with the corresponding results reported in Table 2, LRCS also becomes better by utilizing the target samples as the dictionary.

Finally, our MVDA method outperforms MVDG reported in Table 2. It also achieves the best results on all datasets by incorporating the unlabeled target domain samples when learning the classifiers.

## 6. Conclusion

In this paper, we have proposed a multi-view domain generalization (MVDG) approach for visual recognition, which can effectively fuse multi-view features and simultaneously enhance the domain generalization ability to any unseen target domain. Moreover, we further extend our MVDG approach to a new MVDA approach for domain adaptation by utilizing the target domain data in the training process. The effectiveness of our methods MVDG and MVDA is demonstrated by comprehensive experiments.

## Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. This work was also supported by the Singapore MoE Tier 2 Grant (ARC42/13). This research was also supported by funding from the Faculty of Engineering & Information Technologies, The University of Sydney, under the Faculty Research Cluster Program.

## References

[1] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013.

[2] M. B. Blaschko, C. H. Lampert, and A. Gretton. Semi-supervised laplacian regularization of kernel canonical correlation analysis. In *ECML-PKDD*, 2008.

[3] J. Blitzer, S. Kakade, and D. P. Foster. Domain adaptation with coupled subspaces. In *AISTATS*, 2011.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[6] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *T-PAMI*, 32(5):770–787, 2010.

[7] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIOPT*, 20(4):1956–1982, 2010.

[8] L. Chen, L. Duan, and D. Xu. Event recognition in videos by learning from heterogeneous web sources. In *CVPR*, pages 2666–2673, 2013.

[9] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *ECCV workshop on Consumer Depth Cameras for Computer Vision*, 2012.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[11] Z. Ding and Y. Fu. Low-rank common subspace for multi-view learning. In *ICDM*, 2014.

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[13] L. Duan, I. W. Tsang, and D. Xu. Domain transfer multiple kernel learning. *T-PAMI*, 34(3):465–479, 2012.

[14] L. Duan, D. Xu, and I. W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *T-NNLS*, 23(3):504–518, 2012.

[15] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *T-PAMI*, 34(9):1667–1680, 2012.

[16] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-taylor, and S. Szedmak. Two view learning: SVM-2K, theory and practice. In *NIPS*, 2005.

[17] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[18] B. Gong, K. Grauman, and F. Sha. Reshaping visual datasets for domain adaptation. In *NIPS*, 2013.

[19] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[20] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[21] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[22] M. Hoai and A. Zisserman. Discriminative sub-categorization. In *CVPR*, 2013.

[23] J. Hoffman, K. Saeko, B. Kulis, and T. Darrell. Discovering latent domains for multisource domain adaptation. In *ECCV*, 2012.

[24] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[26] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *T-PAMI*, 36(6):1134–1148, June 2014.

[27] W. Li, L. Niu, and D. Xu. Exploiting privileged information from web data for image categorization. In *ECCV*, pages 437–452, 2014.

[28] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

[29] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of Exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

[30] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[31] L. Niu, W. Li, and D. Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, pages 2774–2783, 2015.

[32] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *T-NN*, 22(2):199–210, 2011.

[33] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[34] V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *ICML workshop on learning with multiple views*, 2005.

[35] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[36] S. Xiao, W. Li, D. Xu, and D. Tao. FaLRR: A fast low rank representation solver. In *CVPR*, pages 4612–4620, 2015.

[37] S. Xiao, M. Tan, and D. Xu. Weighted block-sparse low rank representation for face clustering in videos. In *ECCV*, pages 123–138, 2014.

[38] Z. Xu, W. Li, L. Niu, and D. Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, pages 628–643, 2014.

[39] P. Yang and W. Gao. Multi-view discriminant transfer learning. In *IJCAI*, 2013.

[40] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *ACCV*, 2014.

[41] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence. Multi-view transfer learning with a large margin approach. In *SIGKDD*, 2011.